Scaling and Data Saturation in Protein Language Models

Aviv Spinner [1] **Erika DeBenedictis** [1] **Corey M. Hudson** [1]

Abstract

Data in biology is redundant, noisy, and sparse. How does the type and scale of available **data** impact model performance? In this work, we specifically investigate how protein language models (pLMs) scale with increasing pretraining data. We investigate this relationship by measuring the performance of protein function prediction on a suite of pLMs pretrained on yearly snapshots of UniRef100 from 2011 to 2024. We find no evidence of model saturation on this task: performance improves—but not monotonically—with added data, and this trend differs between unsupervised and supervised experiments. Using a well-characterized β-Lactamase protein from *E. coli*, we find that unsupervised model predictions get better year-over-year, though they do not yet consistently perform better than the supervised baseline. Our results underscore the need for targeted data acquisition and deeper study of data scaling in protein modeling.

1. Introduction

Protein fitness prediction and design is in a period of explosive growth. The successful application of large language models (LLMs) to biological problems has spurred the development of powerful tools. While scaling laws related to compute resources, model parameters, and data quantity have become well-established in the field of natural language processing (NLP) (Kaplan et al., 2020; Hernandez et al., 2021; Hoffmann et al., 2022), analogous studies around **biological data** have not yet been well-characterized. Although the field is beginning to explore scaling in terms of biological systems (Elnaggar et al., 2021; Hesslow et al., 2022; Cheng et al., 2024; Fournier et al., 2024; Li et al., 2024; Nguyen et al., 2024), to our knowledge, there is currently no comprehensive study that investigates how scaling laws relate to data scaling for the task of zero-shot and semi-supervised protein fitness prediction.

1.1. Biological Data

The growth of biological data over the past two decades has been extraordinary. This has enabled the biological machine learning community to answer difficult questions about the interplay of data scale, model performance and how these relate to challenging tasks in biological prediction. Much of this growth in biological data has been in the form of massive sequencing databases such as Uniprot (Consortium, 2024), MGnify (Richardson et al., 2022), OMG (Cornman et al., 2024), and the Big Fantastic Database (BFD)(Jumper et al., 2021). Uniprot is the oldest of these and has experienced considerable growth over its lifetime concomitant with improvements in sequencing. MGnify has also provided considerable growth in sequences, through the incorporation of billions of non-redundant metagenomic assembly sequences. These databases have provided the biological machine learning community with fantastic opportunities to grow the field and expand it into previously impossible questions, evidenced by all of the models trained on these collection of sequences (Jumper et al., 2021; Cornman et al., 2024; Notin et al., 2022; Lin et al., 2023; Madani et al., 2023). Despite the dramatic growth of protein sequence databases, current sequencing efforts capture only a tiny fraction of nature's true protein diversity. Estimates suggest that Earth harbors up to 10^{12} microbial species, most of which remain unsequenced and many of which possess proteins of previously unexplored functions, highlighting how little of the protein universe has been incorporated into AI models (Louca et al., 2019). This gap in sampling and sequencing space represents a potentially fundamental challenge in model sufficiency and protein fitness prediction.

On the other hand, several highly curated data repositories exist that provide annotations and experimental measurements of protein mutational fitness data (Rubin et al., 2025). Resources such as ProteinGym (Notin et al., 2023a) and the TAPE benchmark (Rao et al., 2019) have become standard for evaluating machine learning models on tasks like mutation effect prediction and transfer learning. Testing model performance on experimental data is necessary, and still it only describes a narrow slice of the protein universe and cannot fully capture the breadth, complexity, and noisiness

[*]Equal contribution [1]The Align Foundation. Correspondence to: Aviv Spinner <aviv@alignbio.org>, Corey M. Hudson <corey@alignbio.org>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

of biological data encountered in real-world applications.

1.1.1. Challenges

However, more data does not necessarily equate to better AI model performance. Biological data presents several unique challenges, only some of which are improved by increased data abundance:

- **Redundancy and imbalance**: Overrepresentation of specific families or taxa can bias training and obscure generalization (Ding & Steinhardt, 2024; Poux et al., 2016).
- **Annotation sparsity**: Many sequences lack experimental validation or consistent functional labels (Rauer et al., 2021).
- **Noisy and heterogeneous data sources**: Sequences originate from a mix of high-throughput and manual pipelines, often with varied quality standards (Chorlton, 2024).
- **Functional ambiguity**: Proteins can have multiple or context-dependent functions, making supervised learning difficult (Jeffery, 2023).

In spite of these data challenges, there has been consistent growth in the applications of protein language models (pLMs) across a variety of tasks. However, pLMs are not fully task agnostic. More complex protein tasks (e.g., moonlighting proteins, experimental outcomes, protein fitness and function, etc.) put a greater burden on the model's underlying ability to generalize (Zhou et al., 2024). In practice, increasing the complexity of biological tasks demands more richly labeled datasets.

1.2. Scaling Laws

Scaling laws have helped researchers balance resources between compute, parameters, and training data in order to make optimal models. Teams have measured scaling laws of parameters and compute for protein language models (Elnaggar et al., 2021; Cheng et al., 2024; Fournier et al., 2024; Li et al., 2024; Serrano et al., 2024) as well as published scaling laws experiments as new models are released (Notin et al., 2022; Bhatnagar et al., 2025; Rives et al., 2021).

Despite this, our understanding of biological data scaling laws remains exceptionally limited. While there is consensus that 'more data is better', this heuristic often obscures the nuanced reality that not all data contributes equally to model performance. As previously noted, high sequence redundancy can degrade model performance while strategic data curation can improve outcomes (Cornman et al., 2024). Diminishing returns in performance are frequently observed as datasets grow in size (Gordon et al., 2024).

Recent work has begun to address these questions. The AMPLIFY suite of models (Fournier et al., 2024), trained on UniRef100 snapshots, spanning from 2011-2024, offers a unique window into how model performance changes as the biological sequence pretraining data expands. While many models offer publicly available checkpoints across varying parameter sizes, no other set of models (including ESM, ProGen, and other protein language models) provides a systematic investigation of training data effects with checkpoints released across distinct training data splits. Despite their relatively small parameter counts and single-seed training, AMPLIFY models exhibit competitive performance, remarkable speed, and unique training data splits, enabling a new set of questions around data growth and model generalization to be addressed.

1.3. Our contribution

Here, we first seek to understand the performance of pLMs trained on increasing amounts of unlabeled, publicly-hosted sequence data. We use the suite of AMPLIFY models trained on time-based snapshots of UniRef100 from 2011-2024. Notably, we use Spearman correlation of log-likelihoods of ProteinGym sequences and direct experimental measurements of mutant fitness in ProteinGym as our metric for model performance.

If biological data follows scaling laws similar to those observed in other fields, then model performance should improve predictably as more data is incorporated into the model. In this way, we would project data saturation on the field and provide concrete steps for achieving that. To evaluate the presence or absence of this phenomenon, we use the sequence embeddings from those models in concert with assay data on protein sequences to understand how semi-supervised learning performance changes with both unlabeled training data and labeled training data. In both of these cases, we fail to find the clear hallmarks of scaling laws. From this we infer that for the protein function task, we have not yet reached data saturation.

2. Methods

2.1. Task Datasets

We focus on protein variant effect prediction using the ProteinGym benchmark (Notin et al., 2023a), specifically the substitution datasets from the ProteinGym 1.0 release, with proteins shorter than AMPLIFY's context limit of 2048 amino acids. The `DMS_score` from the resulting 213 datasets is used as the phenotypic label of protein function. We exclude four datasets because of max context lengths of sequences in AMPLIFY: `A0A140D2T1_ZIKV_Sourisseau_2019`; 9576 variants, `BRCA2_HUMAN_Erwood_2022_HEK293T`; 265

variants, POLG_HCVJF_Qi_2014; 1630 variants and POLG_CXB3N_Mattenberger_2021; 15711 variants, in total comprising 1% of ProteinGym.

2.2. Models

2.2.1. Zero-shot

There is only one collection of models, to our knowledge, that share a unified training scheme across many different pretraining datasets. We therefore use the suite of 14 AMPLIFY models (Fournier et al., 2024) trained on yearly releases of UniRef100 from 2011 to 2024.

Sequence Log Probability: To compute the zero-shot "fitness" of a protein sequence, we follow the standard practice of approximation using the log probability from the model. We apply a softmax over the output logits at each position to extract a probability for each of the actual tokens, and sum their logarithms.

To evaluate zero-shot performance, we compare the log-likelihood assigned by each model to the corresponding DMS_score label for each mutant in ProteinGym, using the Spearman correlation coefficient as the evaluation metric, shown in Figure 1.

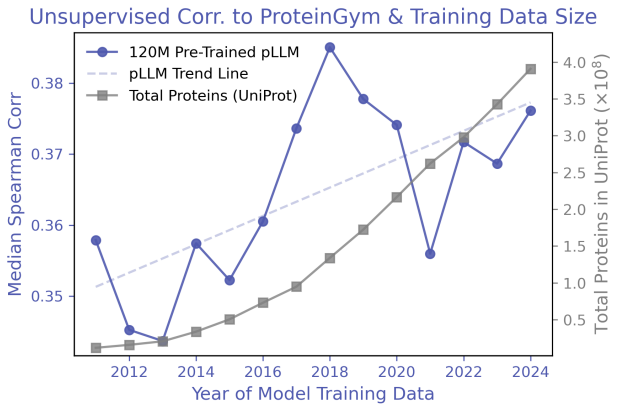

Figure 1. The total number of sequences stored in UniProt from 2011-2024 grows year over year while the correlation of an AMPLIFY model trained on these yearly sources of data bounces between a range of performance, though generally shows an increasing trend.

2.2.2. Supervised

Sequence Embedding: To obtain each protein's sequence embedding for supervised learning tasks, we tokenize the sequence, pass it through an AMPLIFY model, extract the hidden states from the final layer, and compute the mean over the sequence dimension.

We train ridge regression models (Hoerl & Kennard, 1970) using sequence embeddings as input features (x) and DMS fitness values as labels (y) to study scaling in a supervised learning context. Although more sophisticated semi-supervised and supervised architectures exist (Hsu et al., 2022; Notin et al., 2023b; Groth et al., 2024), here we use ridge regression due to its computational efficiency, modularity (i.e., compatibility with diverse inputs and labels), and high interpretability.

Splitting Within Datasets: We implement two versions of the cross-validation schemes introduced in Tranception (Notin et al., 2022) and used in ProteinNPT (Notin et al., 2023b) for benchmarking. In a *Random* split, we randomly partition the data into train/test splits in 10% increments, ranging from 10/90 to 90/10. Each split ratio is replicated five times with different random seeds (Figure 2A). In the *Contiguous* scheme, we divide each protein sequence into five equal-length contiguous segments and train/test on subsets of these contiguous chunks (Figure 2B). In the *Modulo* scheme, we split the positions of the protein into five groups and conduct train/test splits on subsets of these (Figure S5)

Splitting Between Datasets: We would like to understand how information from one assay might help inform future experiments, as this is one of the main motivations for applying machine learning to protein function prediction tasks. ProteinGym contains a protein, β-Lactamase, with four different experiments, performed on nearly identical sets of mutations. In a *structured* split between these different experiments, we train on the earliest experiment from 2012, and test on the remaining three from 2013, 2014, and 2015; train on the earliest two experiments, and test on the remaining two; and train on the earliest three experiments and test on the remaining one dataset. Results are shown in Figure 3.

3. Results and Discussion

We find that model performance does not increase monotonically with the amount of sequence data used for training. The Spearman correlation between experimental data and model-predicted fitness fluctuates year-to-year, showing some decreases even with more training data (Figure 1). For instance, there is a consistent drop in performance between the years of 2018 to 2021, despite an additional billion sequences (and several hundred billion tokens) in UniRef100. One possible interpretation is that between 2014 and 2018, the sum of data added to the model had more relative influence on the model's predictive capabilities of protein function than the data that were added between 2019 and 2021. Overall, the variability in correlation suggests that the model remains sensitive to the specific sequences added or removed at each timepoint, indicating that it has not yet reached data saturation sufficient for robust generalization.

At its best year, this suite of models achieves an average

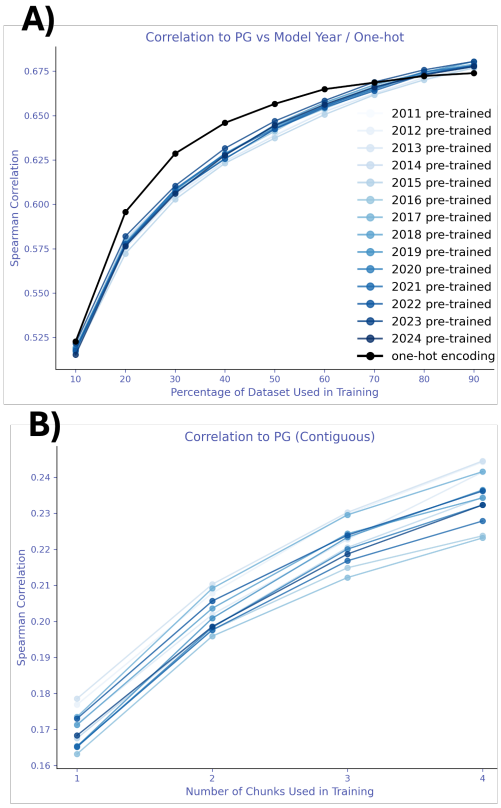

Figure 2. A) Random train/test splits within a dataset training on increasing amounts of labeled data do not reveal large differences from each of the AMPLIFY models. Using one-hot encoding as the sequence representation outperforms model embeddings until a substantial amount of training data is used. B) Contiguous train/test splits within a dataset do not show an obvious trend relating performance to amount of pretraining data seen by the AMPLIFY model.

Spearman correlation of 0.38. When we perform ridge regression, this correlation jumps substantially, leading to the expected boost in performance and continuing to underscore the importance of experimental labels. Even training on 10% of the dataset can increase performance to 0.52, and to around 0.675 when training on 80% of the dataset (Figure 2A).

Both the *Contiguous* (Figure 2B) or *Modulo* (Figure S5) data splitting schemes show that performance improves consistently with the inclusion of more labeled training groups. However, at any given group count, model performance does not show a consistent trend with respect to pre-training year.

Finally, we zoom-in on the β-Lactamase datasets to test the hypothesis that, for this well-characterized protein family with abundant data, scaling laws may begin to emerge. Indeed, we see that the unsupervised predictions of fitness for all four datasets show improvements in correlation with most AMPLIFY models, beginning with a weak correlation of around 0.25 and improving to correlations more than 0.6 (Figure 3). When applying ridge regression in a semi-supervised setup—training on a single dataset and testing on others—we observe a stable correlation around 0.6 across all AMPLIFY models (lightest blue line in Figure 3). Notably, this line intersects the unsupervised curve around 2020, indicating that training on just one experimental dataset can match the zero-shot performance of much larger models trained on a decade of UniRef100 data. As additional datasets are incorporated into training, performance improves monotonically, with two- and three-dataset models outperforming all unsupervised baselines.

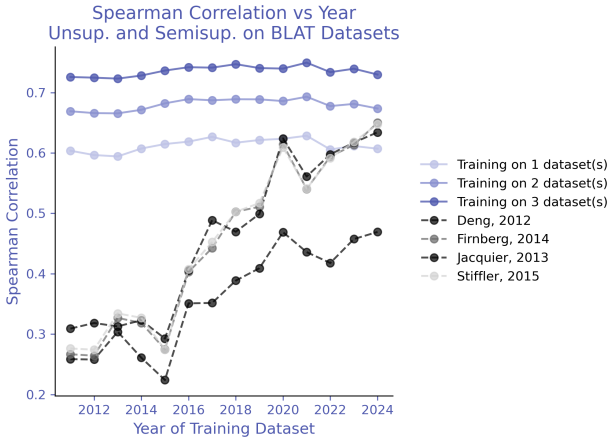

Figure 3. Unsupervised learning is shown with grey traces, and we see that performance does generally get better as the AMPLIFY models train on more UniRef100 data. When doing semi-supervised learning–shown in blue traces, we do not see the same improvement and the lines are quite stable.

3.1. Future Work

We plan on extending this analysis to other families of pLMs as well as exploring other methods for semi-supervised learning. Quantifying the improvement year-to-year on an individual assay may also allow more clear trends to emerge, as they did with the β-Lactamase example. We will also test additional splits between datasets to understand scaling laws in transfer learning. We will train on chosen datasets (i.e. prokaryotic, or a specific protein class, etc) and test on the remaining datasets in ProteinGym. We also hope to randomly split ProteinGym between datasets and train on a small number of randomly selected datasets and test on all others (i.e. train on 10 datasets and test on remaining 210+; train on 100 datasets and test on remaining 110+) to see how performance is balanced between large-scale experimental

data and LLMs.

4. Conclusion

Our findings suggest that for the protein function prediction task, biological data scaling does not yet follow a simple, monotonic trend. Even as pretraining datasets grow, model performance remains sensitive to data composition and has yet to saturate. Experimental labels remain essential for boosting predictive power, and targeted benchmarks like β-Lactamase highlight where scaling benefits begin to emerge. Continued exploration of data scaling—across families, tasks, and learning paradigms—is needed to guide the next generation of biological language models.

Impact Statement

Work presented in this paper impacts the biological machine learning community in understanding the scaling laws underpinning biological data. Because we believe that current public data repositories do not saturate biological data for the task of predicting results of deep mutational scanning experiments, we hope this motivates more scientists to pursue data acquisition and curation. We also hope that the field moves towards quantifying existing data in these repositories & creating better metrics for summarizing complex sequence data.

Acknowledgments

Thank you to the GenBio workshop reviewers and Alan Amin, Pascal Notin, Sarah Gurev, and Lood van Niekerk for helpful comments on the manuscript. We would like to thank the Align team for their support and contributions to this work. This work was supported by The Align Foundation, which receives philanthropic funding in part from Griffin Catalyst.

References

Bhatnagar, A., Jain, S., Beazer, J., Curran, S. C., Hoffnagle, A. M., Ching, K., Martyn, M., Nayfach, S., Ruffolo, J. A., and Madani, A. Scaling unlocks broader generation and deeper functional understanding of proteins. *bioRxiv*, pp. 2025–04, 2025.

Cheng, X., Chen, B., Li, P., Gong, J., Tang, J., and Song, L. Training compute-optimal protein language models. *Advances in Neural Information Processing Systems*, 37: 69386–69418, 2024.

Chorlton, S. D. Ten common issues with reference sequence databases and how to mitigate them. *Frontiers in Bioinformatics*, 4, 2024. doi: 10.3389/fbinf.2024.1278228. URL https://doi.org/10.3389/fbinf.2024.1278228. Link.

Consortium, T. U. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1): D609–D617, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

Cornman, A., West-Roberts, J., Camargo, A. P., Roux, S., Beracochea, M., Mirdita, M., Ovchinnikov, S., and Hwang, Y. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, 2024. doi: 10.1101/2024.08.14.607850. URL https://www.biorxiv.org/content/early/2024/08/17/2024.08.14.607850.

Ding, F. and Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life. *BioRxiv*, pp. 2024–03, 2024.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Fournier, Q., Vernon, R. M., van der Sloot, A., Schulz, B., Chandar, S., and Langmead, C. J. Protein language models: is scaling necessary? *bioRxiv*, pp. 2024–09, 2024.

Gordon, C., Lu, A. X., and Abbeel, P. Protein language model fitness is a matter of preference. *bioRxiv*, pp. 2024–10, 2024.

Groth, P. M., Kerrn, M., Olsen, L., Salomon, J., and Boomsma, W. Kermut: Composite kernel regression for protein variant effects. *Advances in Neural Information Processing Systems*, 37:29514–29565, 2024.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022.

Jeffery, C. J. Current successes and remaining challenges in protein function prediction. *Frontiers in Bioinformatics*, 3:1222182, 2023. doi: 10.3389/fbinf.2023.1222182. Link.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K., and Lu, A. X. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pp. 2024–02, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. A census-based estimate of earth's bacterial and archaeal diversity. *PLoS biology*, 17(2):e3000106, 2019.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106, 2023.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024.

Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A. N., Marks, D., and Gal, Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023a.

Notin, P., Weitzman, R., Marks, D., and Gal, Y. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023b.

Poux, S., Magrane, M., Bateman, A., Belda, E., Boeckmann, B., Boutet, E., Breuza, L., Bridge, A., Coudert, E., Esperet, E., Famiglietti, M. L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lieberherr, D., Lombardot, T., Martin, M. J., Masson, P., Natale, D. A., Pedruzzi, I., Roechert, B., Schneider, M., Sigrist, C. J. A., Stutz, A., Sundaram, S., Tognolli, M., Bougueleret, L., Xenarios, I., Bairoch, A., Redaschi, N., and Consortium, U. Expert curation in uniprotkb: a case study on dealing with conflicting and erroneous data. *Database*, 2016:baw139, 2016. doi: 10.1093/database/baw139. URL https://academic.oup.com/database/article/doi/10.1093/database/baw139/2742069.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Rauer, C., Sen, N., Waman, V. P., Abbasian, M., and Orengo, C. A. Computational approaches to predict protein functional families and functional sites. *Current Opinion in Structural Biology*, 70:108–122, 2021. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2021.05.012. URL https://www.sciencedirect.com/science/article/pii/S0959440X21000816. Biophysical Methods.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L., Curtis, T., Escobar-Zepeda, A., Gurbich, T., Kale, V., Korobeynikov, A., Raj, S., Rogers, A., Sakharova, E., Sanchez, S., Wilkinson, D., and Finn, R. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51 (D1):D753–D759, 12 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1080. URL https://doi.org/10.1093/nar/gkac1080.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Rubin, A. F., Stone, J., Bianchi, A. H., Capodanno, B. J., Da, E. Y., Dias, M., Esposito, D., Frazer, J., Fu, Y., Grindstaff, S. B., et al. Mavedb 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biology*, 26(1):13, 2025.

Serrano, Y., Ciudad, Á., and Molina, A. Are protein language models compute optimal? *arXiv preprint arXiv:2406.07249*, 2024.

Zhou, Z., Zhang, L., Yu, Y., Wu, B., Li, M., Hong, L., and Tan, P. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15(1):5566, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49798-6. URL https://doi.org/10.1038/s41467-024-49798-6.

A. Appendix

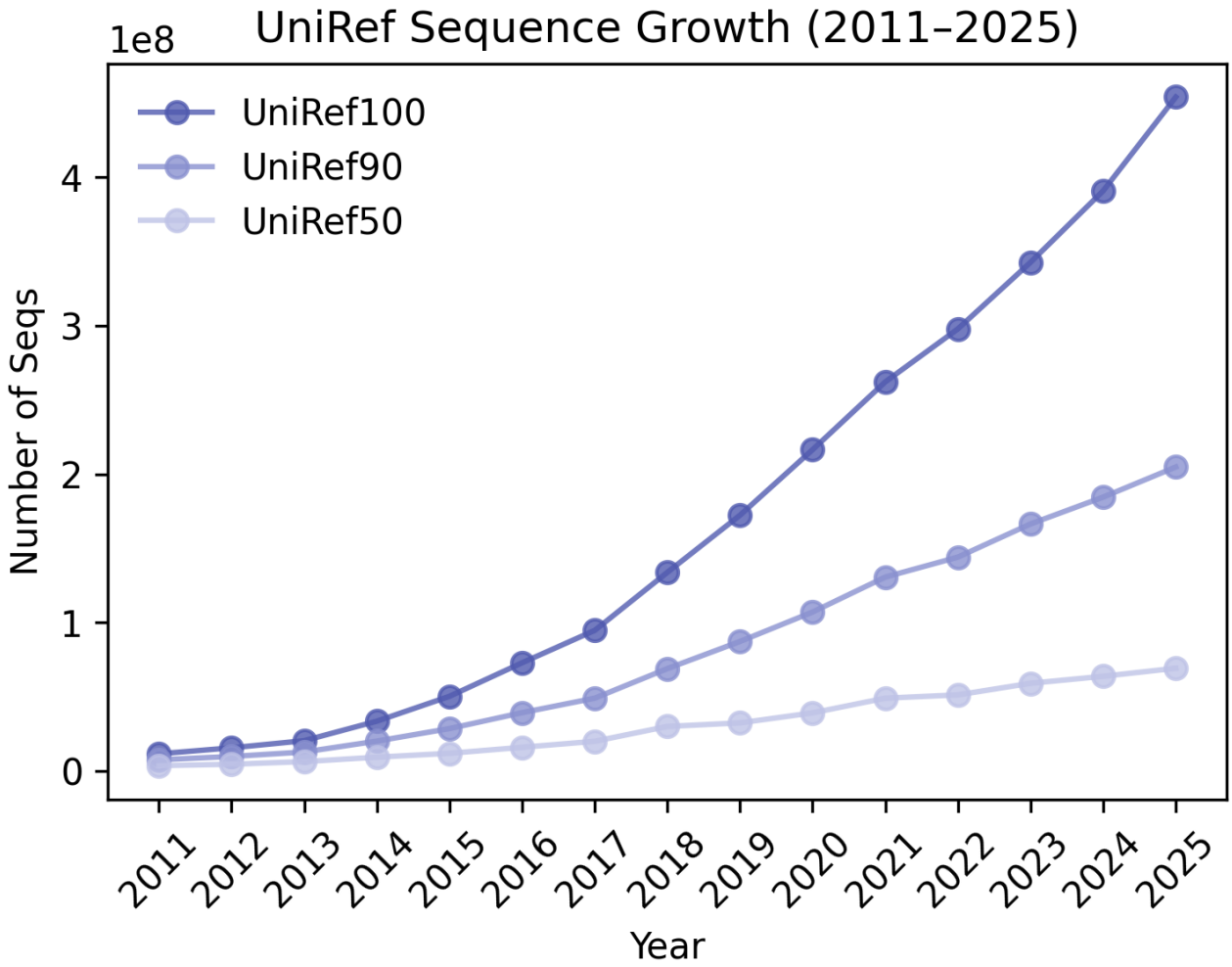

Figure S1. UniRef100, UniRef90, and UniRef50 grow year over year from 2011 to 2025. UniRef100 grows faster than UniRef90 which grows faster than UniRef50, demonstrating that we are collecting more sequences that are more similar to each other.

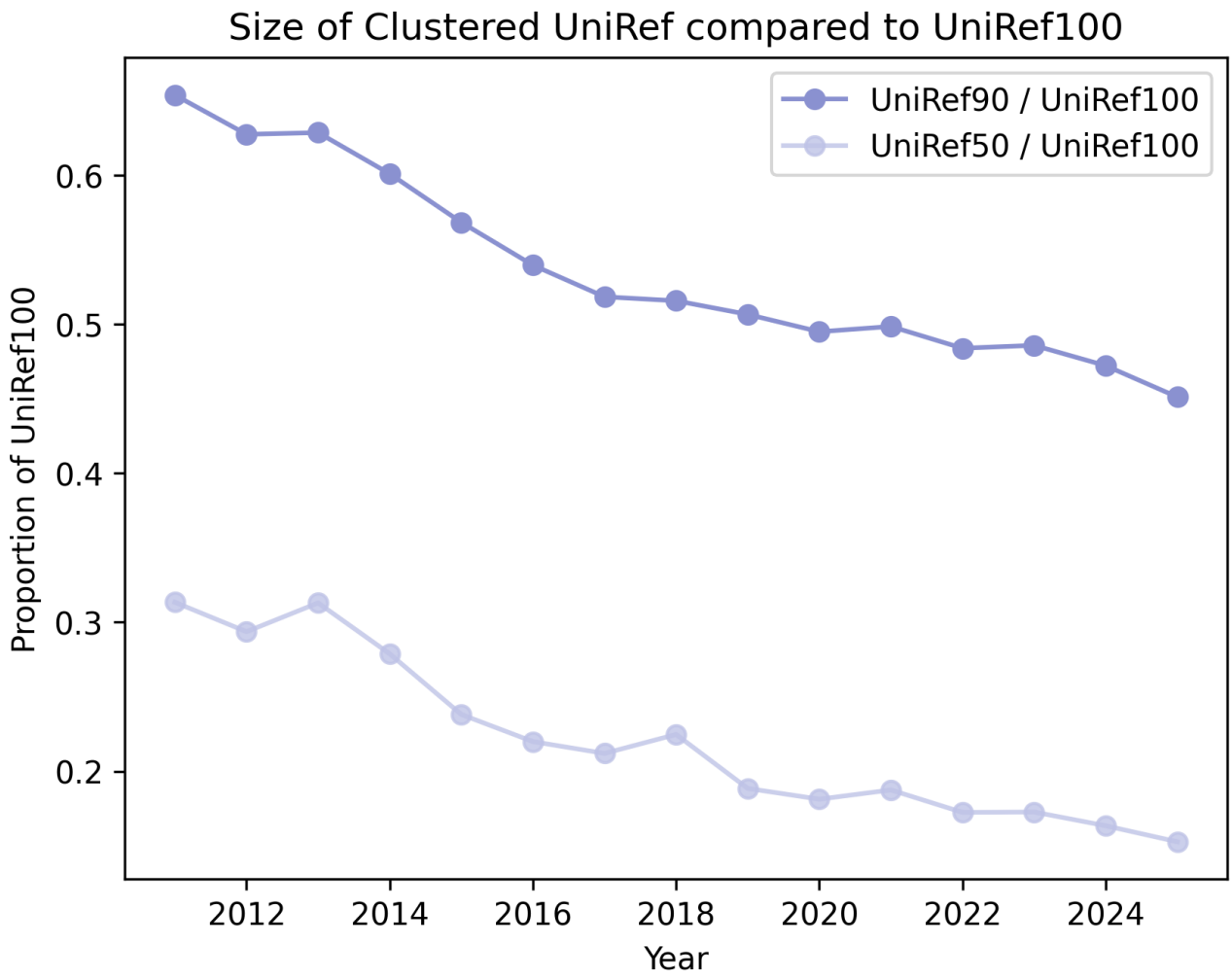

Figure S2. The ratio of UniRef100 with either UniRef90 or UniRef50 shrinks year over year. If every single sequence in UniRef100 was more than 90% or 50% different than every other sequence, we would see a flat line at 0.9 and 0.5 respectively. However, we see that both of these trends start lower than that and shrink each year. Therefore we can infer that the data added continues to be less and less diverse.

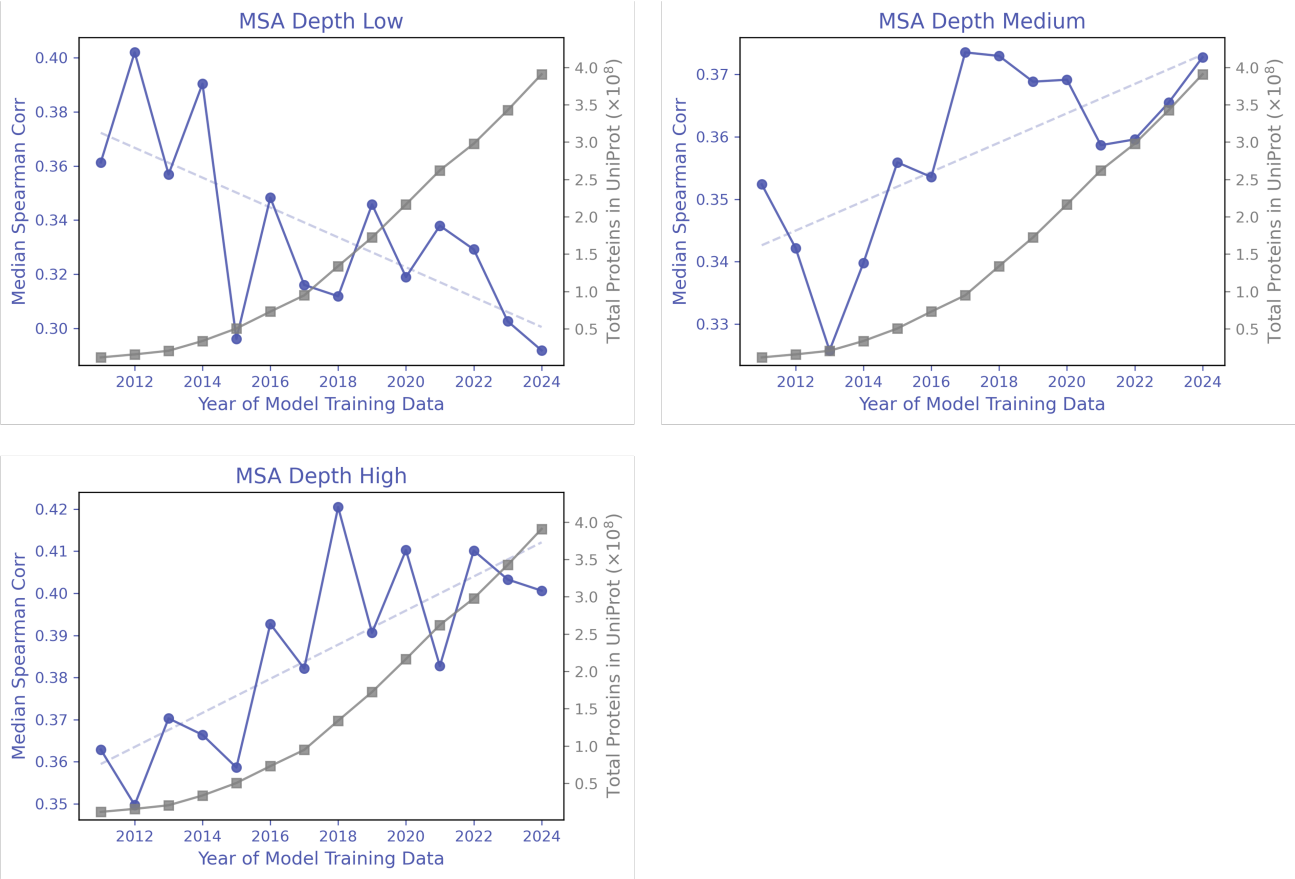

Figure S3. Figure S3 replicates Figure 1, partitioned by partitioned by MSA depth, as represented as Neff/L from ProteinGym. Proteins with Low MSA depth generally get worse with later timepoints while proteins with Medium and High MSA depth generally improve.

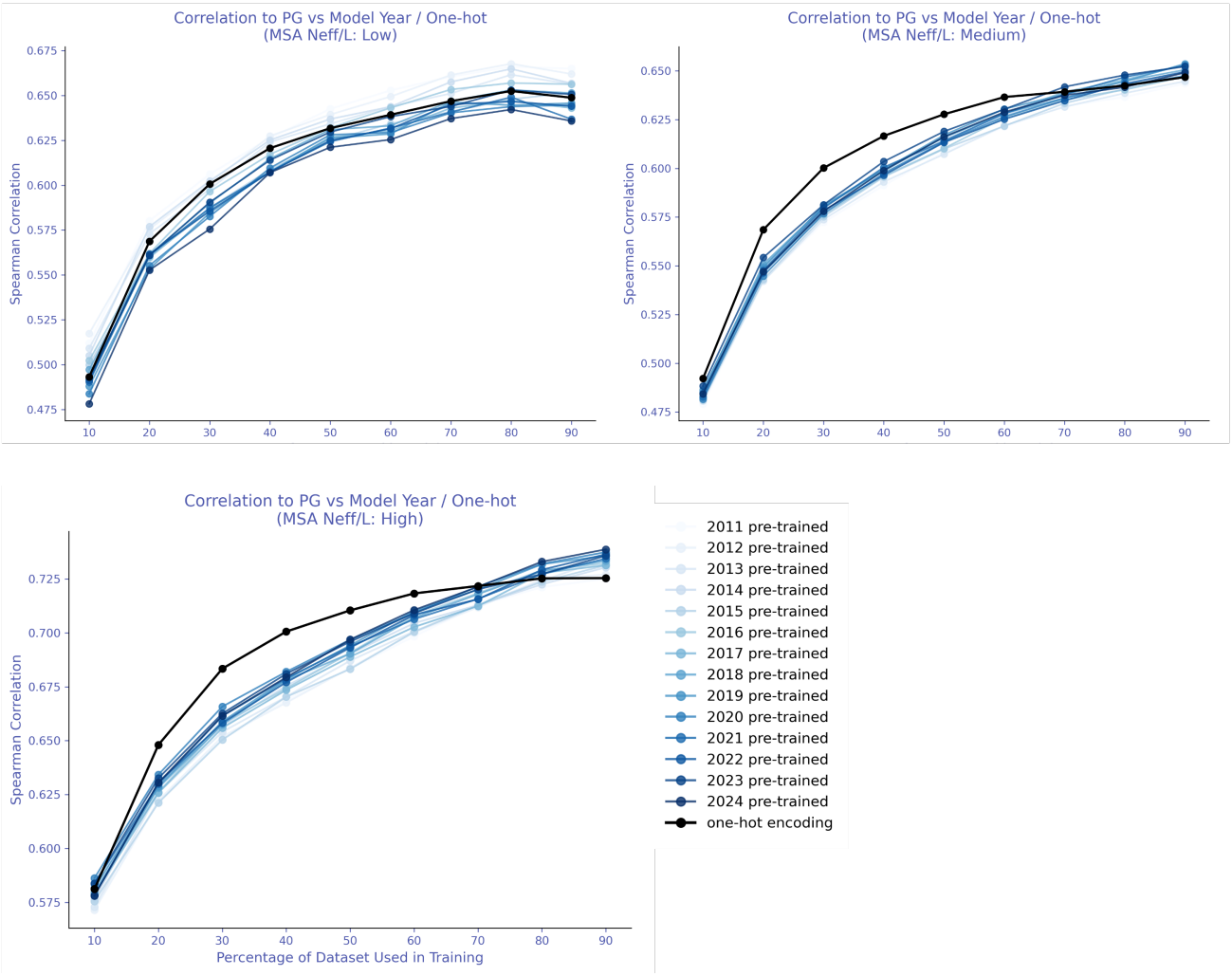

Figure S4. Figure S4 replicates Figure 2A, partitioned by by MSA depth, as represented as Neff/L from ProteinGym. Proteins with Low MSA depth exhibit better performance with models trained on earlier timepoints of UniRef100. Whereas proteins with Medium and High MSA depth do not show as large of a range.

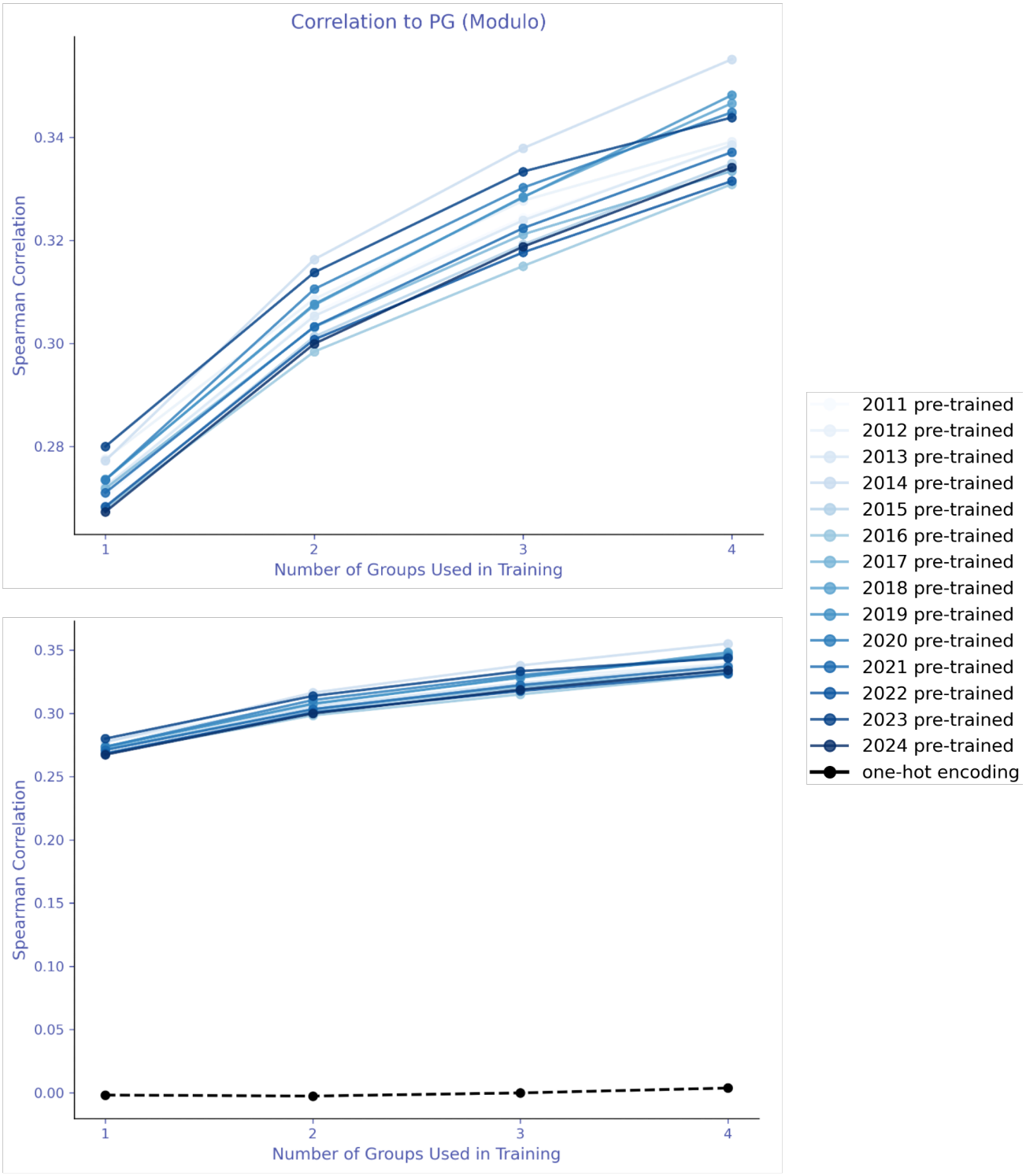

Figure S5. This is a similar plot as Figure 2B but with the Modulo train/test split. The top plot shows the results without one-hot plotted and the bottom plot shows the results with one-hot plotted. One-hot encodings provide almost no information to the model when splitting in this non-random way and the correlation hovers around 0 regardless of how much labeled data is used.

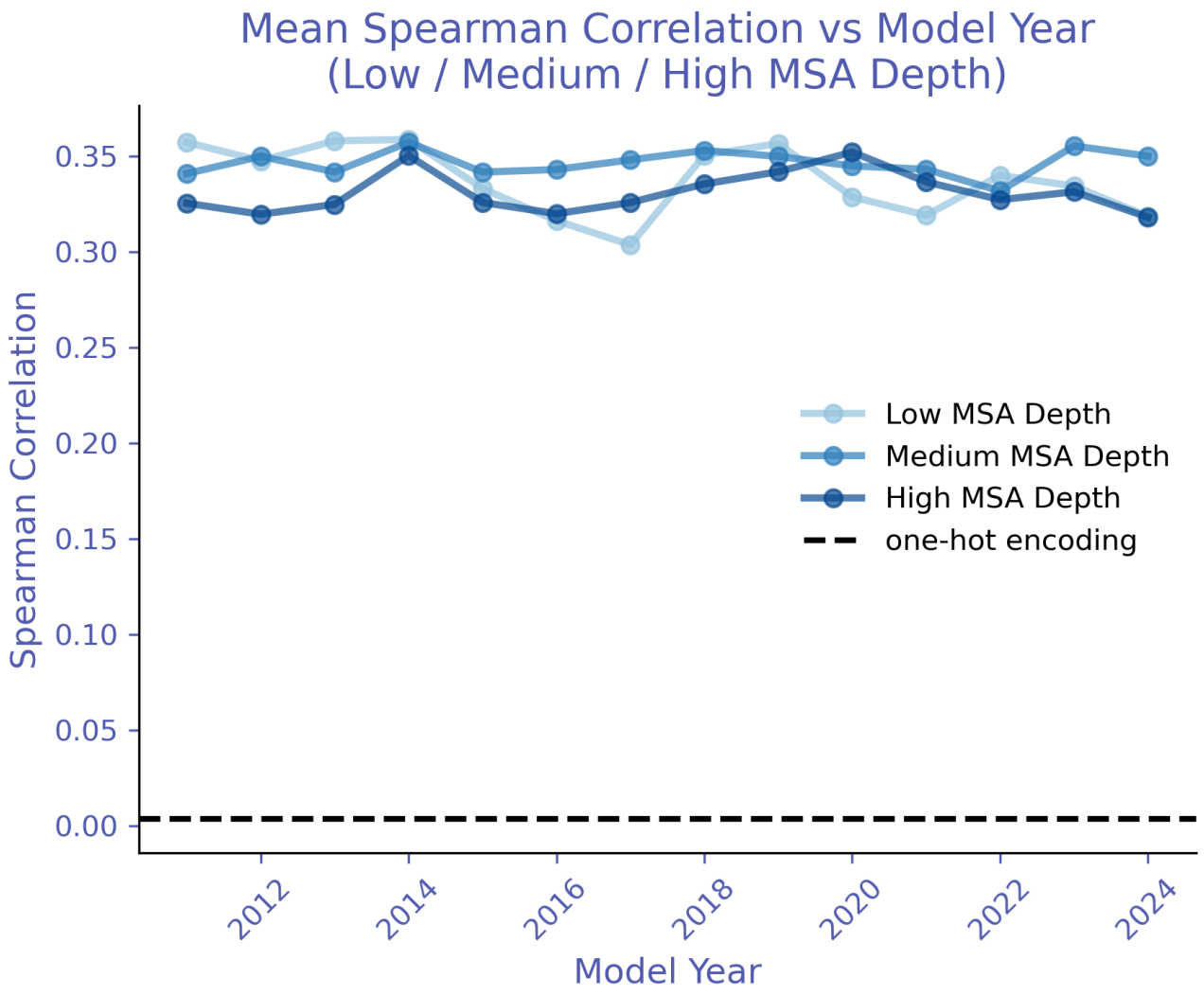

Figure S6. When looking at the Modulo train/test splits by MSA depth, we see that the trend is similar between them and relatively flat.

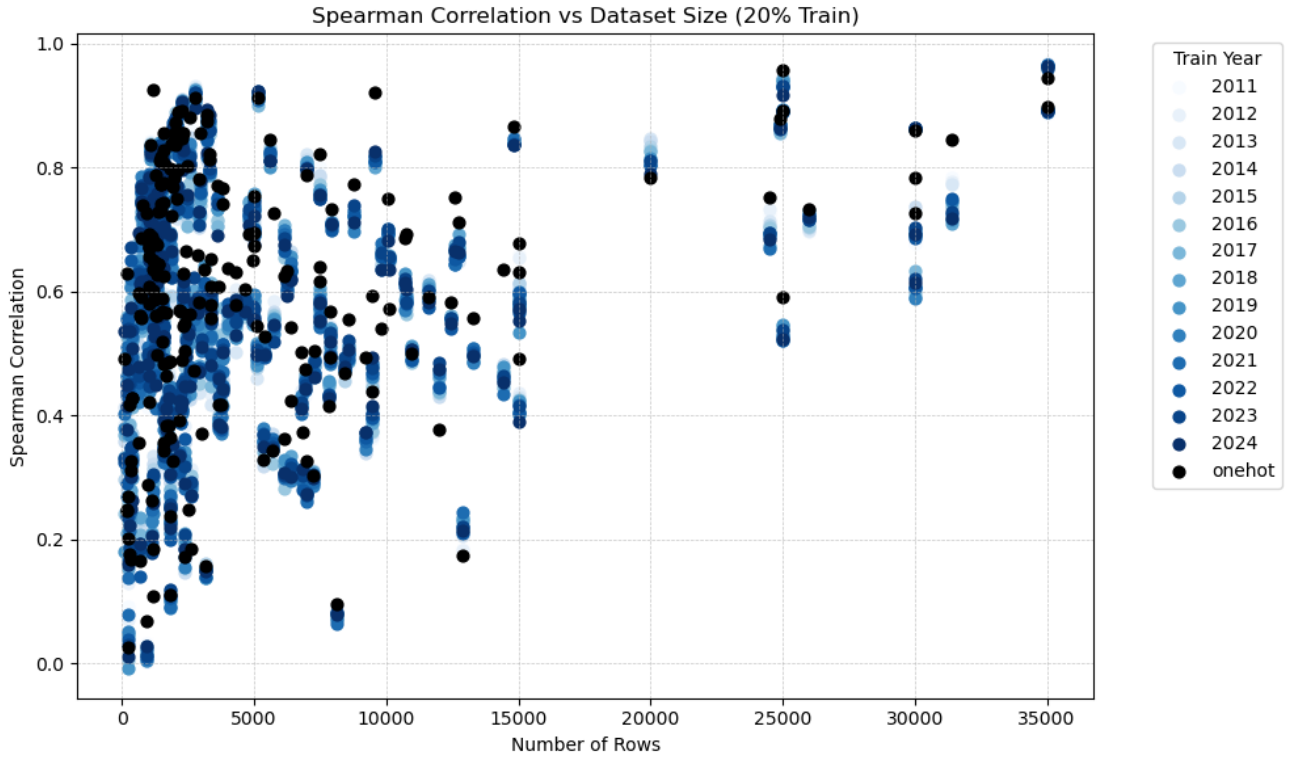

Figure S7. The number of mutants included in the dataset impacts the spread of spearman correlation between the model predictions and the experimental values. With relatively smaller dateset sizes, the spearman correlations range from 0 to close to 1. Whereas with larger dataset sizes (15,000 mutants measured), the model accuracy is always above 0.5. This is only for training on a random split of 20% of the data and we see a similar trend across all random data splits.

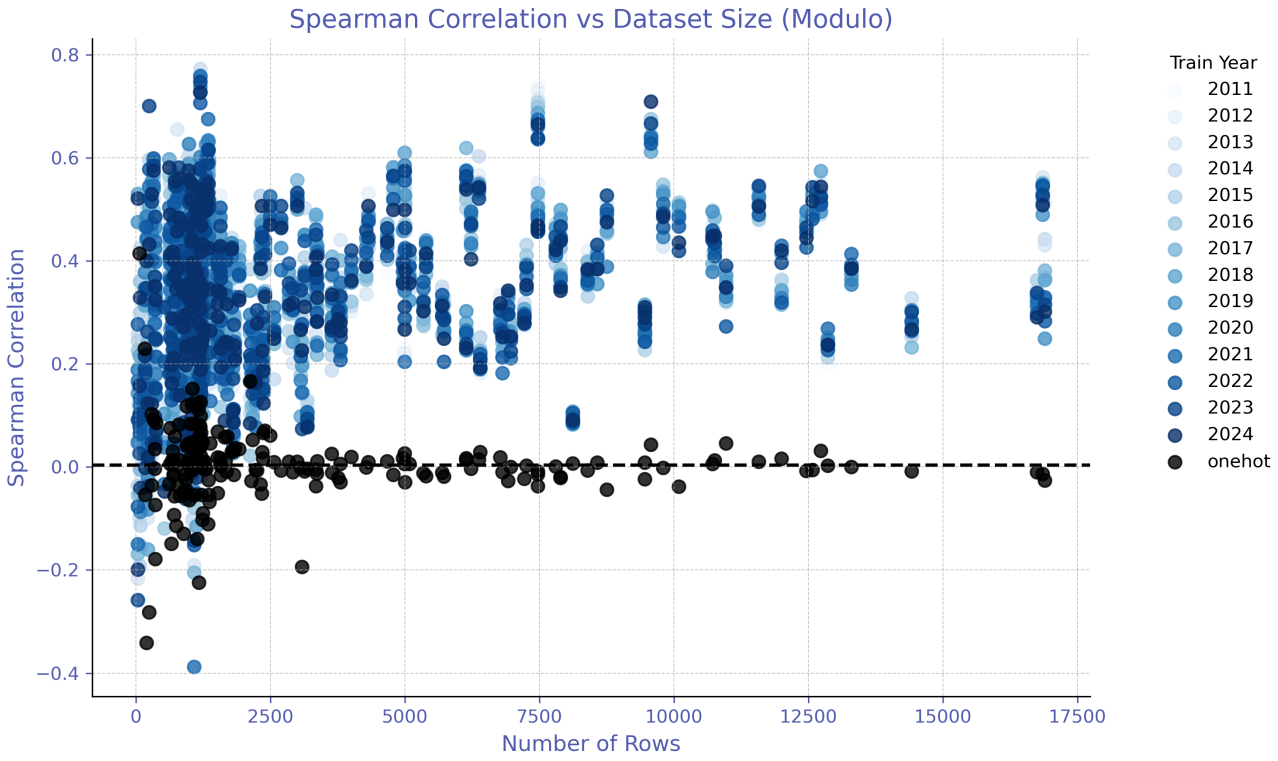

Figure S8. Identical Figure S7 but with the Modulo train/test split instead of random. Generally we see a similar trend of more spread out data at lower number of mutations (rows). One-hot encoding, however, does look incredibly different and hovers around 0, with the mean correlation shown in the dashed line.